CLAIMS

1. A method for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, said method comprising the step of performing a linear regression analysis using data from a dataset of matching genotypes and phenotypes,

   wherein the log fold resistance, pFR, of each HIV strain is modelled as the sum of all the individual resistance contributions for each of the mutations or combinations of mutations that occur in HIV according to the following equation;

$$pFR = \beta_A M_A + \beta_B M_B + \beta_n M_n + \cdots + \beta_Z M_Z + \varepsilon$$

   wherein each individual resistance contribution is calculated by multiplying a mutation factor, M_A , M_B , ..., M_Z, for each mutation or combination of mutations by a resistance coefficient β_A, β_B, ..., β_Z;

   wherein for a combination of mutations, the mutation factor M_n represents the co-occurrence of one mutation with other one or more mutations and the coefficient β_n represents the synergy or antagonism between the one mutation with the other one or more mutations;

   wherein the mutation factor assigned to each mutation or combination of mutations reflects the degree to which that mutation or combination of mutations is present in the HIV strain and, if present, to which degree the mutation is present in a mixture;

   wherein each resistance coefficient reflects the contribution of the mutation or combination of mutations to the fold resistance exhibited by the strain;

   and wherein the error term ε, represents the difference between the modelled prediction and the experimentally determined measurement.

2. A method according to claim 1, wherein correlations are removed from the dataset for correlated mutations where not all correlated mutations contribute to the drug resistance phenotype, using an algorithm to track the change in pFR for each mutation as the effects of individual mutations or combinations of mutations are removed from the dataset.

3. A method according to claim 2, wherein the algorithm performs the following steps:

   a) calculate average pFR for all mutations with a sufficient count in the database to be significant;

b) determine the extremes (maximum, minimum), and select the mutation with the pFR furthest away from the global average;

c) remove all virus strains that have the selected mutation from the dataset and reiterate from step a);

d) stop when the selected mutation in step b) has an average pFR that approximates to the global average;

such that removing virus strains with a certain resistance causing mutation results in an increase of the average pFR for correlating mutations, which thus have a higher average pFR.

4. A method according to any one of claims 1-3, wherein the algorithm performs the following steps:

a) calculate correlation coefficient between all mutations with a sufficient count in the database and the pFR;

b) determine the extremes (maximum, minimum), and select the mutation with the highest absolute value of correlation coefficient;

c) calculate a linear model for the pFR with the selected mutation(s) (from step b), all previous iterations);

d) take the residue;

e) calculate correlation coefficient between all mutations with a sufficient count in the database and the residue;

f) determine the extremes (maximum, minimum), and select the mutation with the highest absolute value of correlation coefficient;

g) calculate a linear model for the pFR with the selected mutation(s) (from step f), all previous iterations); and

h) reiterate from step d);

i) stop when the selected mutation in step g) has a correlation coefficient that approximates to zero.

5. A method according to any one of the preceding claims, wherein censored values in the genotype / phenotype database are replaced by a maximum likelihood estimation.

6. A method according to claim 5, wherein for each iteration of the linear regression, the following steps are performed until the predictions converge:

   a) Calculate a linear regression model without censored values;

   b) Use the phenotypic measured value V_0 as if the censor was " = ", e.g. when a result is expressed as –log *FR* < 4, we will treat V_0 as –log *FR* =4;

   c) Look at the prediction P from the model and apply either:

      When case '<'-censor:

      - $P < V_0 - 0.798\ \sigma$ (center of gravity of half Gaussian distribution)
        - Remove value from training data for next iteration
      - $V_0 - 0.798\ \sigma \leq P < V_0$
        - Use $V' = V_0 - 0.798\ \sigma$ for next iteration
      - $V_0 \leq P$
        - Use V' centre of gravity of tail (<V) of a normal distribution N (P, σ) as value for next iteration.

      When case '>'-censor:

      - $P > V_0 + 0.798\ \sigma$ (center of gravity of half Gaussian distribution)
        - Remove value from training data for next iteration
      - $V_0 + 0.798\ \sigma \geq P > V_0$
        - Use $V' = V_0 - 0.798\ \sigma$ for next iteration
      - $V_0 \geq P$
        - Use V' centre of gravity of tail (>V) of a normal distribution N (P, σ) as value for next iteration.

   d) Calculate a linear regression model and for the censored values in the linear regression model, either remove the data-point from the training set, or use V' instead of the censored phenotypes measurement, as described in step c);

   e) Re-iterate from step b) until prediction converges.

7. A method of identifying a mutation that effects the degree of drug resistance exhibited by an HIV strain using a method according to any one of the preceding claims.

8. A method according to claim 1 wherein the contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain is calculated, said method comprising the steps of:

a) obtaining a genetic sequence of said HIV strain,

b) identifying the pattern of mutations in said genetic sequence, wherein said mutations are associated with resistance or susceptibility to drug therapy, and

c) calculating the fold resistance of the HIV strain as compared to the wild type HIV strain by performing a linear regression analysis according to claim 1.

9. A method according to claim 8, which incorporates a method according to any one of claims 1-7.

10. A method according to any one of the preceding claims, wherein in the case of small datasets for particular mutations or combinations of mutations, the method is applied recursively to the set of virus strains that exhibit those particular mutations or combinations of mutations.

11. A diagnostic method for optimising a drug therapy in a patient, comprising performing a method according to any one claims 7-9 for each drug or combination of drugs being considered to obtaining a series of drug resistance phenotypes and therefore assess the effect of the plurality of drugs or drug combinations on the predicted fold resistance exhibited by the HIV strain with which the patient is infected and selecting the drug or drug combination for which the HIV strain is predicted to have the lowest fold resistance.

12. A method according to any one of claims 7-9 and 11, wherein the resistance coefficient for each mutation is calculated using a method according to any one of claims 1-6.

13. Use of a method according to any one of claims 1-6 for assessing the efficiency of a patient's therapy or for evaluating or optimizing a therapy.

14. A diagnostic system for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by an HIV strain, said system comprising:

a) means for obtaining a genetic sequence of said HIV strain;

b) means for identifying the mutation pattern in said genetic sequence as compared to wild type HIV;

c) means for predicting the fold resistance exhibited by the HIV strain using any one of the methods of claims 1-12.

15. A computer apparatus or computer-based system adapted to perform the method of any one of the claims 1-12.

16. A computer program product for use in conjunction with a computer, said computer program comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising a module that is configured so that upon receiving a request to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, it performs a method according to any one of claims 1-12.